Chapter 1
Introduction

1.1. Background

Speech recognition is a technology which has close connections with computer science, signal processing, voice linguistics and intelligent systems. The communications between humans and computers are wider and deeper and communication functions by using mouse, keyboard and touch screen cannot satisfy the quick, accurate and efficient interchange of information. How to send information in a more natural, more efficient and quicker way has become an urgent question.

From technology research to daily life, computers are involved in every aspect of people's daily life. Computers are used to accomplish many tasks. Considering this situation, intelligent communication between computers and humans, human-computer interaction, becomes one of the most important research fields.

Speech is one of the natural forms of human communication. Since childhood people can express themselves by speech, recognizing others by distinguishing their voices and under-standing others by their speech. People are very good at speaker and speech recognition. The human brain uses neurons and synapses, modified with experience and provides a distributed form of associative memory. Motivated by this, speaker and speech recognition systems have been developed. Speaker recognition is the technology of letting a machine distinguishes different speakers from each other. Depending on the different speakers different actions are implemented.

Speech recognition is the technology of letting a machine understand human speech and, according to the meaning of the speech, implement the intention of the human.

1.2. History of Speech Recognition

Speech recognition technology is one of the most popular and potential technologies. Generally speech recognition has changed the computer into an "intelligent" device. Speech is the most natural communication medium. With the development of computer science and speech processing, translation between different languages will be the active part of speech processing research. The design of a natural speech database, feature extraction of speech, using speech material to do acoustical model research, speech recognition arithmetic research, language translation and conversation processing research will be the speech technology's hotspot direction. In the middle of the 20th century, speech recognition was created as a new subject. At that time the AT&T Bell laboratory implemented the first speech recognition system, which could recognize ten English numbers, the "Audry System". In 1997 IBM developed a Chinese Via Voice speech recognition system and many other companies developed speech recognition systems e.g. Speech Works has the automatic phone speech recognition system, Microsoft and Speech Works cooperated to integrate speech recognition technology in Office XP, and developed the Microsoft Speech SDK. All this indicates that speech recognition technology has become more and more well versed and it will be one of the most important technology trends in the future information field.

In the 1960s the applications of computer sciences promoted the development of speech recognition. At this time the most important achievements were: the generation of Dynamic Programming (DP) and Linear Predictive (LP) technologies.

In the 1970s LP technology had developed further, the Dynamic Time Warping (DTW) technology was well developed, especially Vector Quantization (VQ) and Hidden Markov Model (HMM) had been put forward. In practice, systems for people with isolated word recognition based on LP and DTW technologies were achieved. In the 1980s the main achievement is the continuous speech recognition. Many algorithms for continuous speech recognition have been developed e.g in Bell laboratory, Myers, Rabiner and Lee's hierarchical construction algorithm and frame synchronization. At the same time the research direction changed from template matching to statistical modeling techniques. HMM became the main speech recognition technology. In 1988, Carnegie Mellon University (CMU) using VQ/HMM methods implemented

the SPHINX system, which could recognize continuous speech for 997 words, independently of the speaker. This is the first system for speaker independent, large vocabulary and continuous speech. It pioneered a new stage of speech recognition.

In the 1990s speech recognition went through a rapid development, in some fields speech recognition has become mature. As the multimedia time came, there was an urgent demand to implement speech recognition systems in practice. Another main trend was that: speech recognition started to combine with other fields. In the beginning of the 90s there were researches about speech recognition and natural speech processing combined. In the middle of the 90s speech recognition was combined with machine translation technology and the development of translation between different languages started.

1.3. Speech Recognition

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format.

1.3.1. Classification of speech recognition systems

According to the application area, speech recognition systems can be divided into speaker identification systems and speaker verification systems.

1.3.1.1. Speaker identification

Speaker identification is the process of determining which registered speaker gave the given utterance. In the speaker identification system the test utterance is compared and scored with the registered speakers and the one whose module matches the best is selected. Based on whether test utterances are allowed to come from any unknown identity, speaker identification can be divided into two types:

Closed set identification. In this process the system knows that the test utterance belongs to one of the registered speaker already, so the system will force itself to give an identity to this test utterance in any case.

Open set identification. In this process the system does not know if the test utterance belongs to the registered speakers or not. So if the system does not find any register speaker who is matched with the test, the system will reject to assign the utterance to any of the speakers.

The performance of speaker identification depends on two main factors:

Text-dependent or text-independent. Text-dependent requires the speakers to say exactly the given or the same speech. The system knows beforehand what the speakers will say and so it is easier and quicker to make the decision. Although higher performance could be achieved by knowing the utterance beforehand, such systems are more prone to cheating. Text-independent does not require the speakers to utter the same speech, the speaker can be recognized by any speech. Here the system does not know what the speaker will say, a more exible system will be needed. Speaker number and confusability. With more speakers, a larger database will be needed and also longer time for training and scoring. Another point is that when the test utterances are similar in pronunciation or when text-dependent speakers' utterances are similar, a more exible and powerful system is needed to complete this task. System design. With different feature extraction methods, different neural network algorithms, different systems can be designed and different performance can be reached.

1.3.1.2. Speaker verification

Speaker verification is the process of accepting or rejecting the identity claim of a speaker. The system extracts parameters from the input speech signal to represent vocal characteristics and uses these information to build representative speaker models. When a test utterance comes with a claim, it tests its parameters under the claimed speaker's model, and calculates a similarity score. The decision is made depending on this score. If the score is below the threshold then the system will reject the test utterance, otherwise the system will accept the test utterance.

The performance of speaker verification depends on three main factors:

Text-dependent or text-independent-Text-dependent requires the speaker to say exactly the given or the same speech. As technology develops people can imitate others' voice by using some software, so this system's security needs to be considered. Text independent does not require the speaker to utter the same speech, the speaker can be recognized by any speech. For example the system generates random speech, the test speaker reads it and the system makes the decision. This is popular in the application of speaker verification.

Registered utterance-This is the training samples and the basis of calculating similarity score, it has great effect on the performance of the system. On the other hand, the amount of the registered utterance and the duration of the registered utterance will make the decision different.

System design-The method of calculating similarity score, the error resistance, system algorithms all give great effects on the result. Acoustic variability: This is a big challenge towards achieving high performance.

1.3.2. Factors affecting speech recognition performance

In speech recognition systems, different systems have different characters and designs according to the task requirements. But in most cases few factors are the same. These factors play a significant roles for the systems state of accuracy: Vocabulary size and confusability According to how much vocabulary can be recognized, speech recognition can be divided into three kinds of different scales vocabulary speech recognition:

- Small scale vocabulary speech recognition
- Medium scale vocabulary speech recognition
- Large scale vocabulary speech recognition

Small-scale can identify less than 100 vocabulary while medium-scale can identify more than 100 vocabulary and large-scale can identify more than 1000 vocabulary.

The situation that always happens is that when people are talking, the listener misunderstands what the speaker says. For the human brain, which is very good at recognition, this still happens quite often. This is mostly because of the large scale vocabulary and the confusability. A general rule is that it is easy to discriminate among a small set of words. As the size of vocabulary differs, it is different to discriminate among a different set of words. Moreover, if the vocabulary contains confusable words, this will make the task more difficult.

Speaker dependence: This system is used by an exact speaker. So the system needs to recognize different speeches from one speaker.

Speaker independence: This system can be used by any speakers. In this system, the features become tuned to the speaker that it was trained on, and these features tend to be highly speaker-specific. Speech recognition as mentioned before the goal is to recognize what is being said irrespective of who is speaking. Different speakers have different voice, tone, frequency etc, so different features can be extracted although for the same speech. With the speaker-specific features, bigger challenges need to be conquered in order to get a good performance. Isolated, discontinuous, or continuous speech. As we all know, a kid can recognize a separate word earlier than an entire sentence. It is the same situation with a speech recognition system. One question is always considered first: What kind of speech needs to be recognized? Is it a word or a sentence? So the speech that needs to be recognized can be divided into: Isolated speech (single words) e.g identifying ten single numbers from 0 to 9, place names, control commands or Chinese syllables. Discontinuous speech (full sentences) consists of words that are separated by silence. The ultimate purpose of speech recognition is to make the computer understand natural language. The

biggest character of natural language is continuous. Continuous speech recognition is the most difficult task in speech recognition. For example, continuous speech recognition is needed in dictation, translation machines or human-computer speech conversations. Read or spontaneous speech. The way the speech was given is also an important factor that influences the performance of the system. Normal speech can be given in many ways, generally it can be divided into:

. Read speech: Speakers read and are recorded from prepared scripts

. Spontaneous speech: The utterance is spontaneous. Spontaneous speech is vastly more difficult, because it tends to be peppered, false starts, incomplete sentences, stuttering, coughing, and laughter. Moreover, the vocabulary is essentially unlimited, so the system must be able to deal intelligently with unknown words. Other factors affecting both speaker and speech recognition

Other factors that affect the level of difficulty in speech recognition, and therefore also the reliability, are for instance:

- Variability in speakers : Gender, speed, regional and social dialects, speaking style, emotional, physical states
- Variability in environments: background noise, reverberation.
- Variability in transmission channels and microphones
- Variability in expertise: Speech physiology, acoustic phonetics, digital signal processing's, statistical pattern recognition

1.4 Time Plan

Table No. 1.4 Time Plan

	Start (Date)	Aug	Sept	Oct	Jan	Feb	Mar	Apr
Selection of Topic	26/07							
Literature Survey	26/07							
Abstract	25/08							
Implementation I	27/08							
Report I (End of Sem VII)	01/09							
Implementation II	01/01							
Report II	28/02							
Final Report (Proof Reading and Binding)	23/04							